REGRESSION MODELS: LM & GLM

University of Zurich UZH

http://r-bayesian-networks.org/
gilles.kratzer@math.uzh.ch
sonja.hartnack@access.uzh.ch

GILLES KRATZER, APPLIED STATISTICS GROUP, UZH

SONJA HARTNACK, VETSUISSE, UZH

ECVPH WORKSHOP, ZURICH 7-9 MAY 2019

Motivational example

```
# load required packages
library(car)


head(Prestige)
```

	education	income	women	prestige	census	type
gov.administrators	13.11	12351	11.16	68.8	1113	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
accountants	12.77	9271	15.70	63.4	1171	prof
purchasing.officers	11.42	8865	9.11	56.8	1175	prof
chemists	14.62	8403	11.68	73.5	2111	prof
physicists	15.64	11030	5.13	77.6	2113	prof

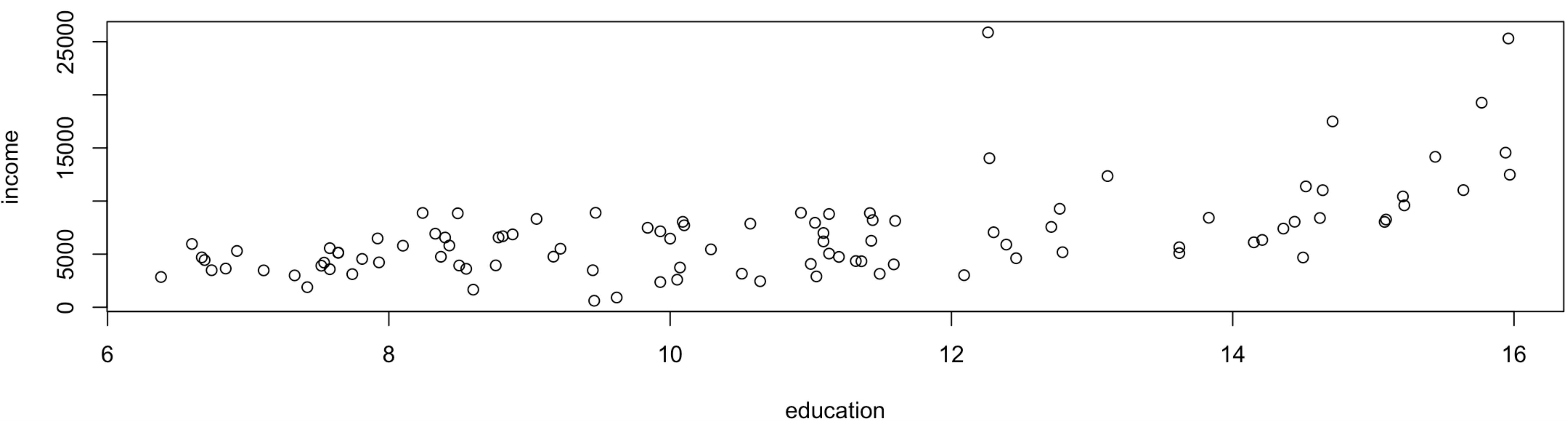
Motivational example

```
> str(Prestige)
'data.frame':   102 obs. of  6 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

Motivational example

```
# Subset the data to keep only income and education
newdata = Prestige[,c(1:2)]
summary(newdata)

plot(newdata)
```



Motivational example

```
summary(lm(formula = income~education,
           data = newdata))
```

```
Call:
lm(formula = income ~ education, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-5493.2 -2433.8   -41.9  1491.5 17713.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2853.6     1407.0  -2.028   0.0452 *
education      898.8      127.0   7.075 2.08e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3483 on 100 degrees of freedom
Multiple R-squared:  0.3336,    Adjusted R-squared:  0.3269
F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10
```

Simple linear model: definition

In univariate linear regression a dependant variable is explained linearly through a single independent variable:

Parameters (unknown)

Measurement error

$$y_i = \alpha + \beta_1 x_i + \epsilon_i, \text{ with } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

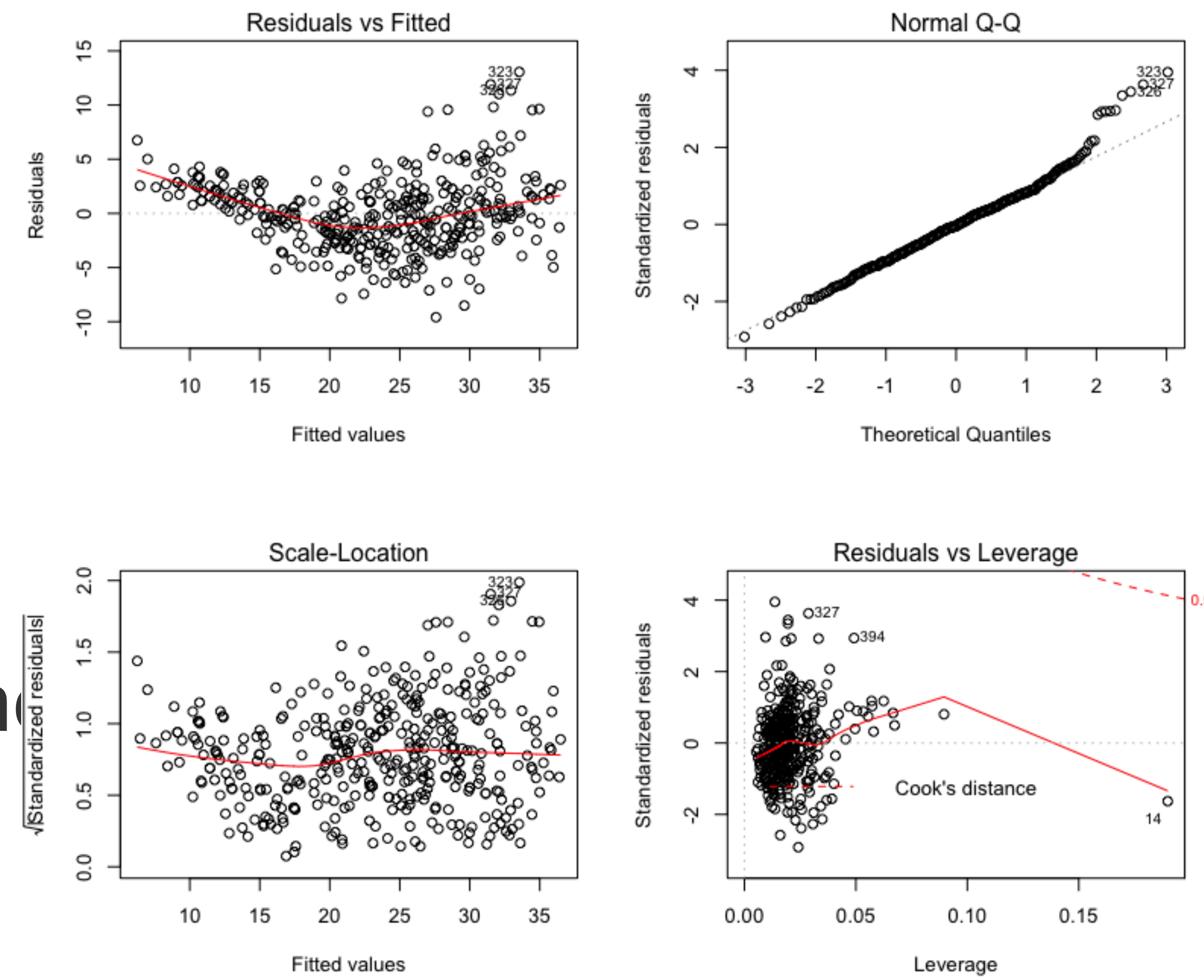
Dependent variable, measured values, observations

Independent variable, measured values, observations

Errors follows a normal distribution 0 mean, constant variance

Linear model: assumptions

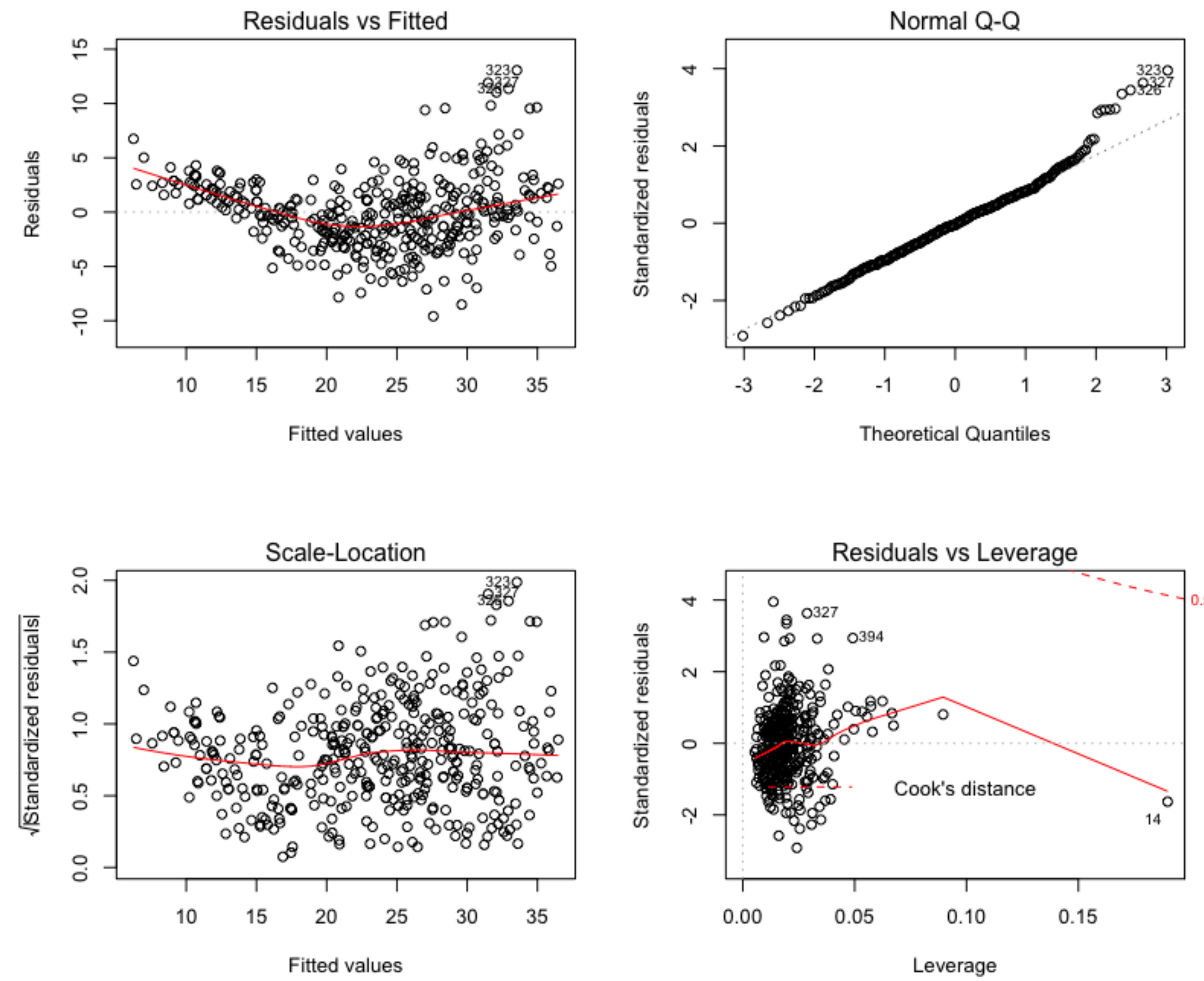
1. Linear relationship
2. Multivariate normality
3. Homoscedasticity
4. No or little multicollinearity
5. No autocorrelation



In R 4 diagnostic plots are return

- Residual vs Fitted Values
- Normal Q-Q Plot
- Scale Location Plot
- Residuals vs Leverage Plot

Linear model: assumptions



1. Linear relationship
2. Multivariate normality
3. Homoscedasticity
4. No or little multicollinearity
5. No autocorrelation

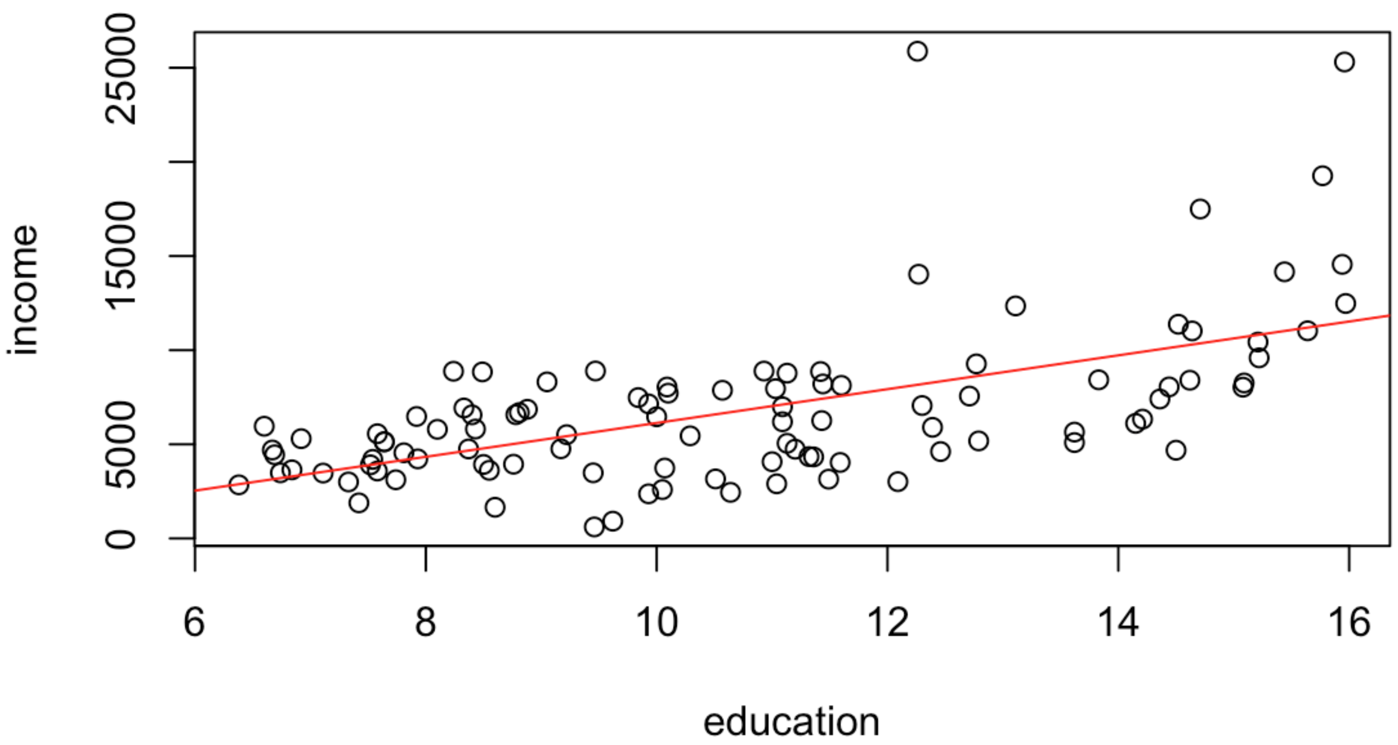
In R 4 diagnostic plots are returned:

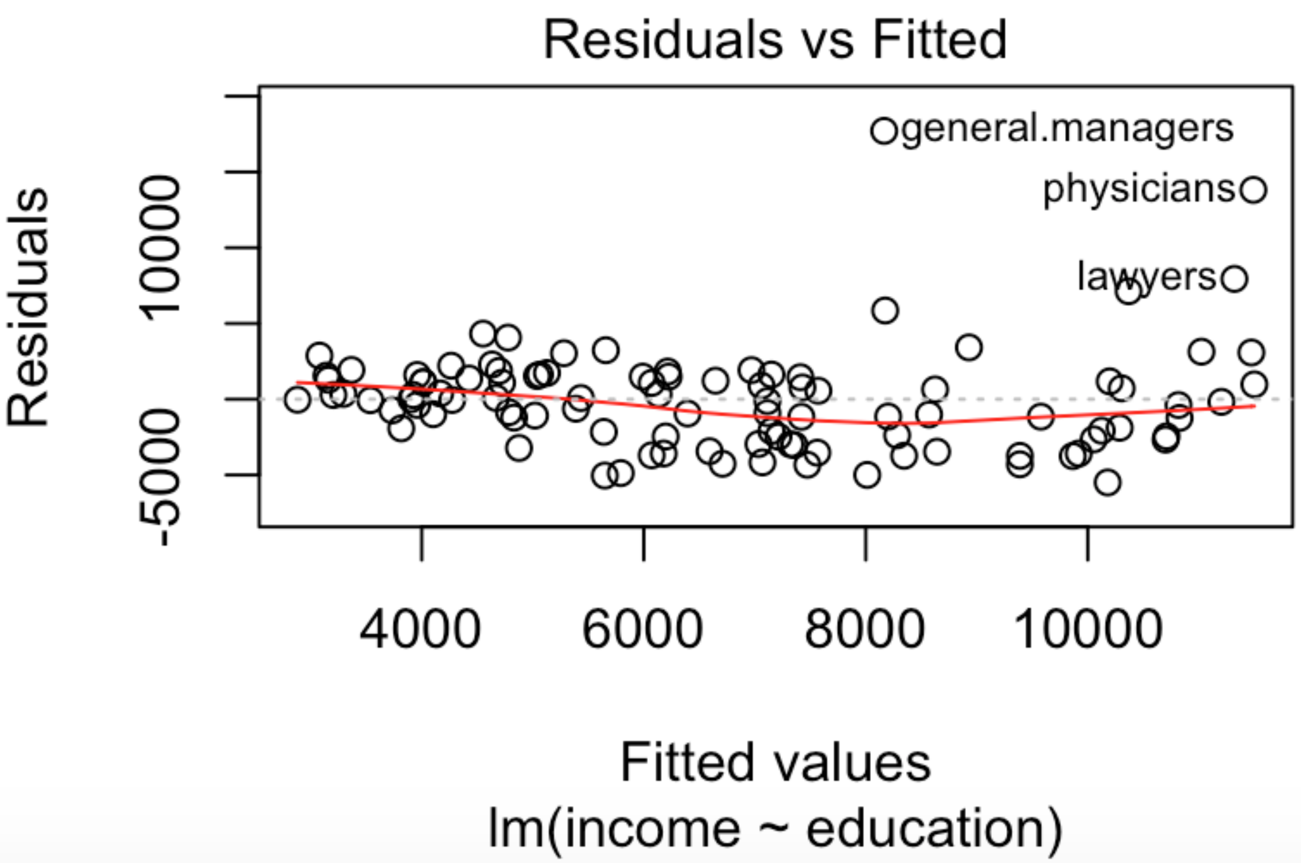
- Residual vs Fitted Values (1)
- Normal Q-Q Plot (2)
- Scale Location Plot (3)
- Residuals vs Leverage Plot (outliers)

Linear model: assumptions

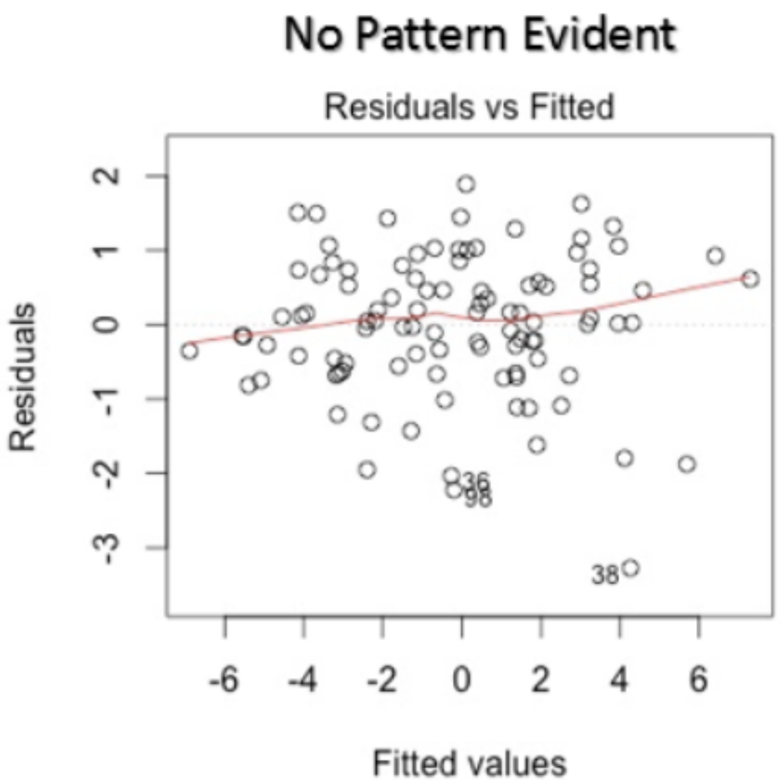
- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation

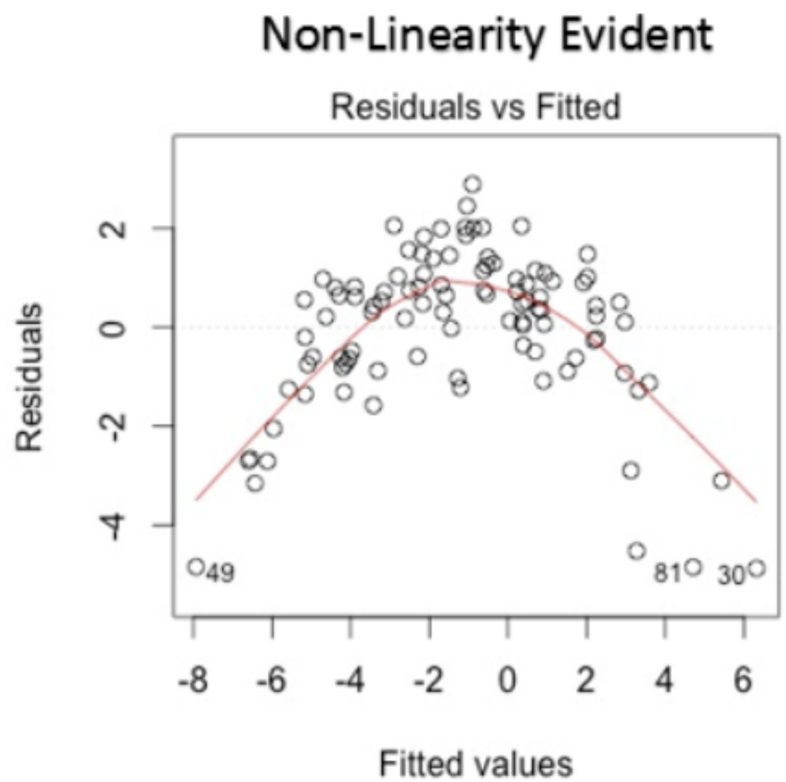
Linear model: linear relationship





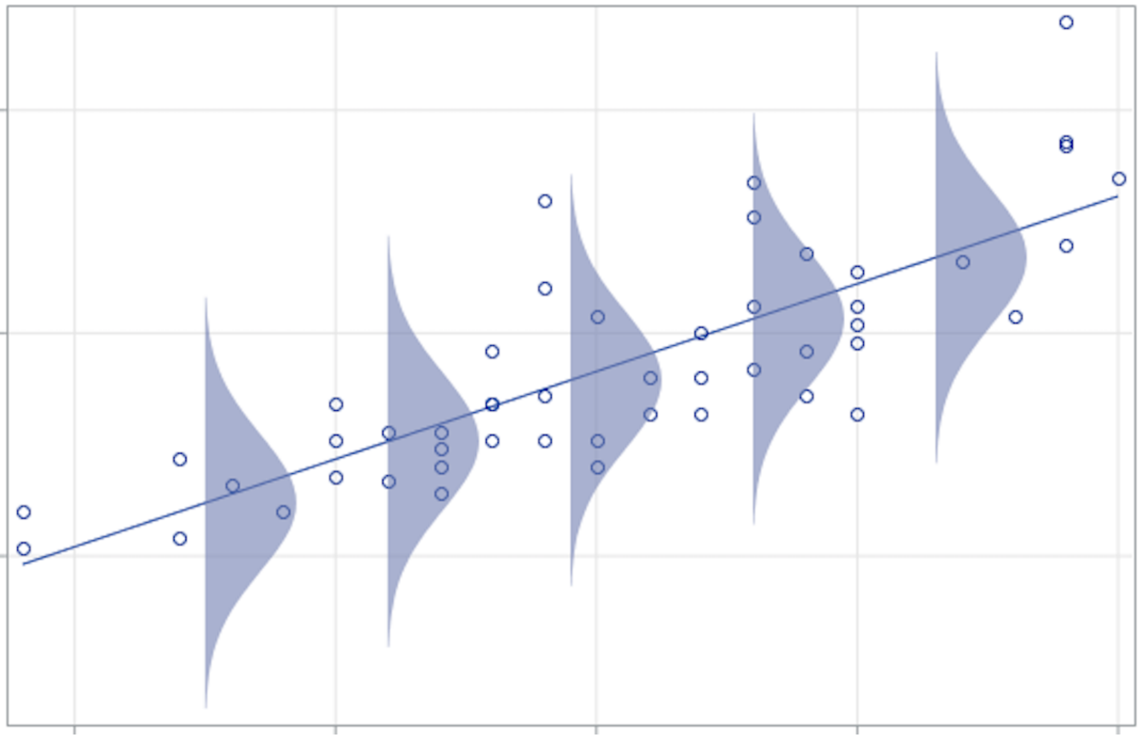
Linear model: linear relationship





Linear model: linear relationship

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity



Error term should be normally distributed

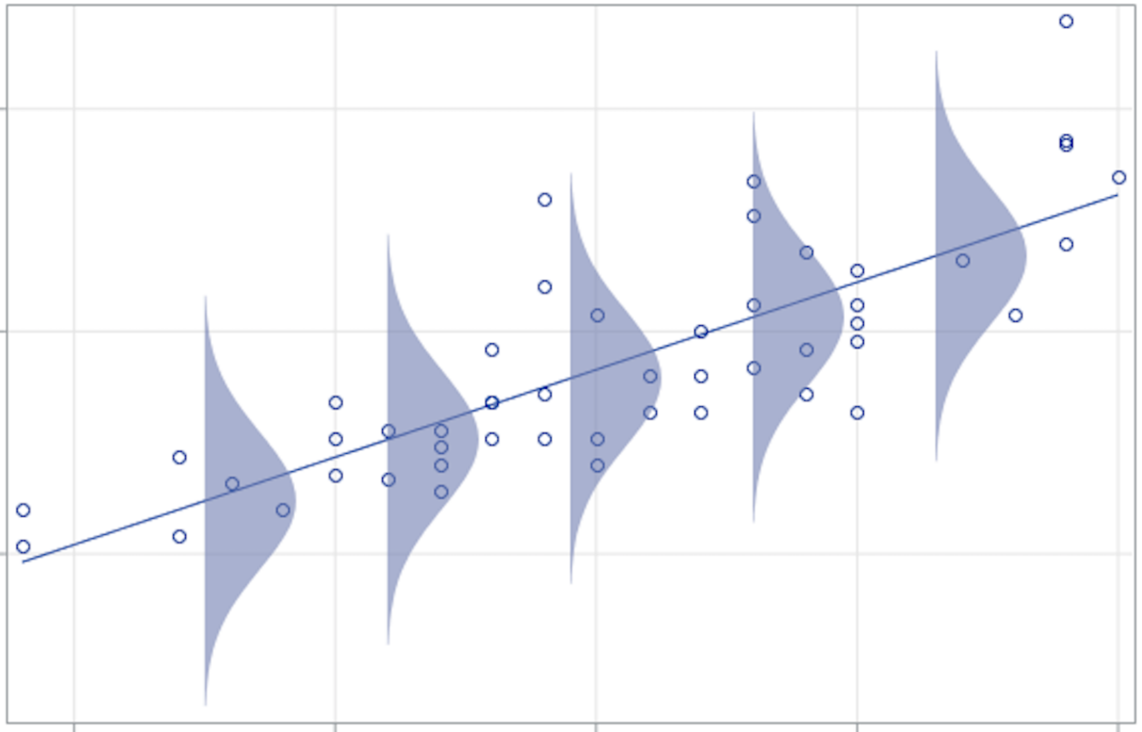
$$\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}})$$

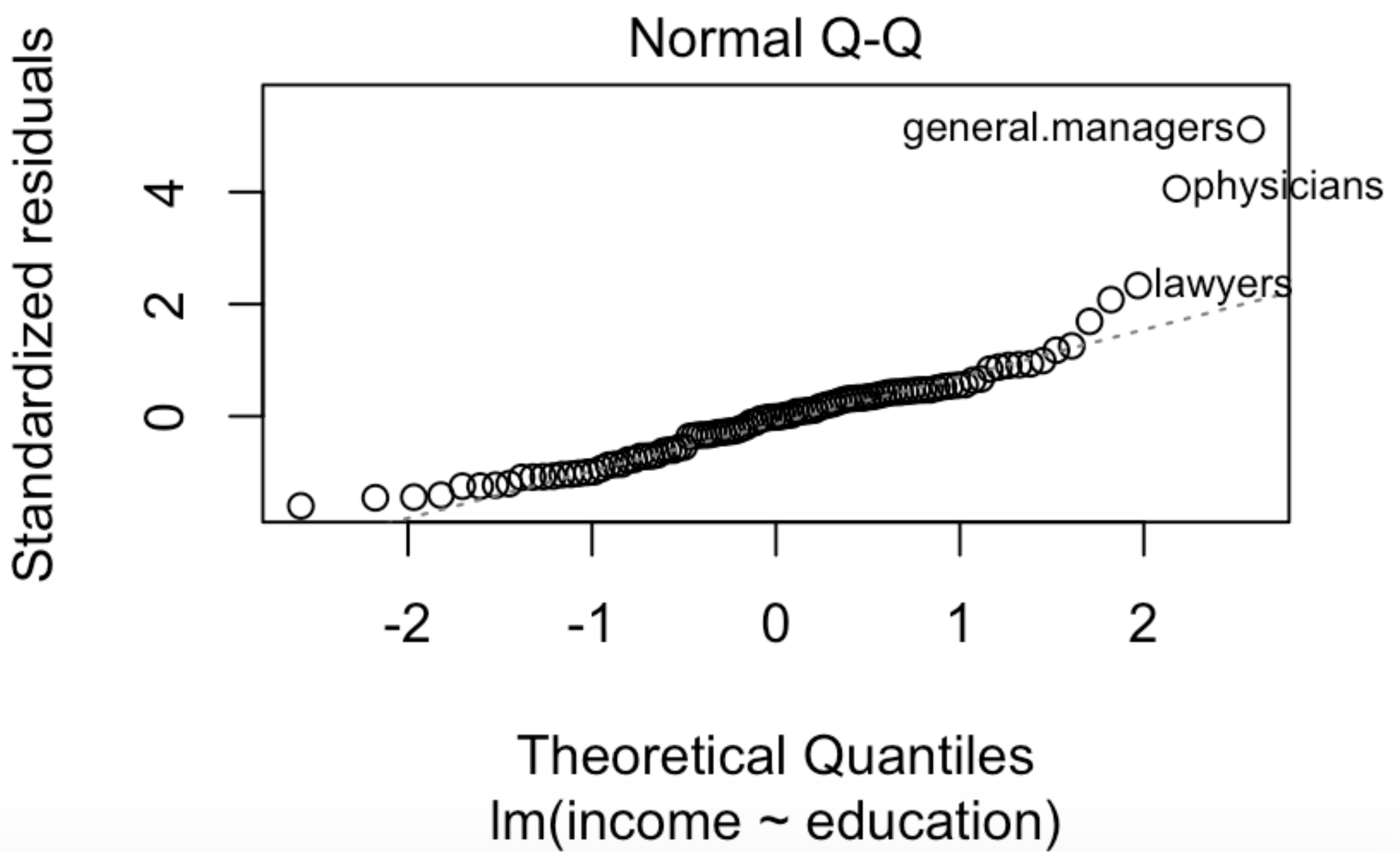
Conditional response should be normally distributed

$$(\mathbf{y} \mid \mathbf{x}) \sim \mathbf{N}(\mathbf{X}^{\mathbf{T}}\beta, \sigma^{\mathbf{2}})$$

Linear model: Normality

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity

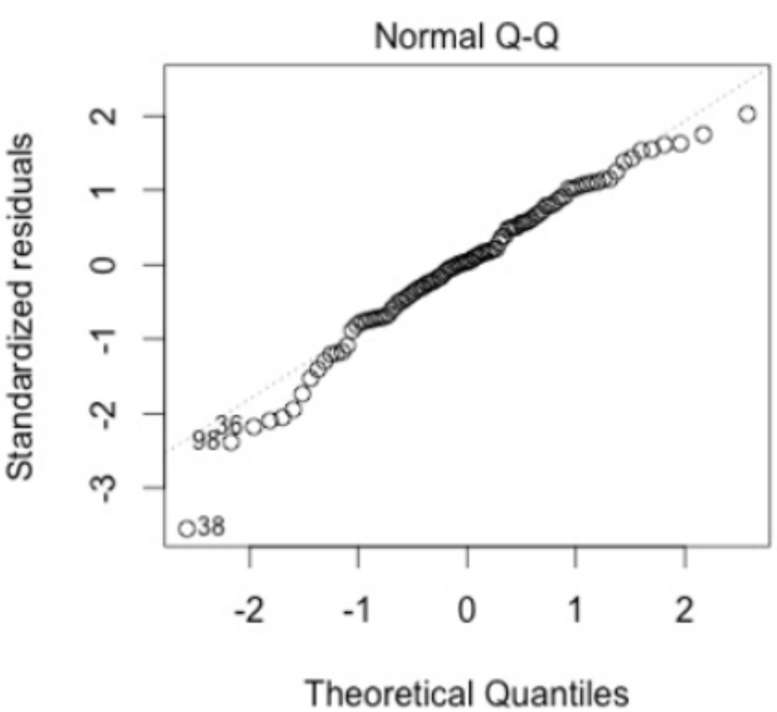




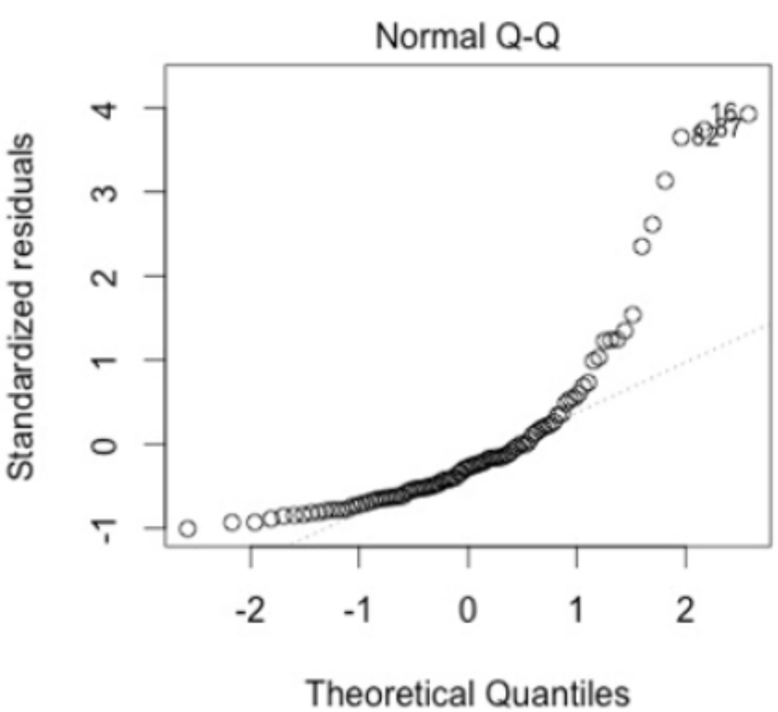
Linear model: Normality

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation

Normal Distribution Evident

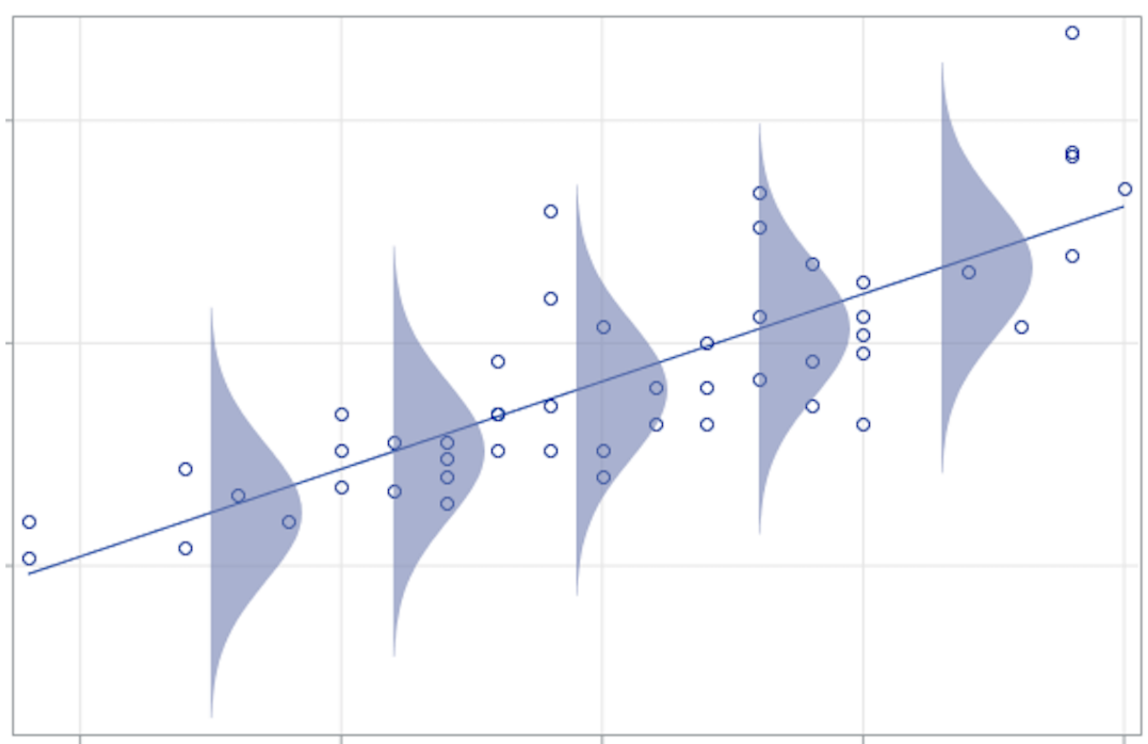


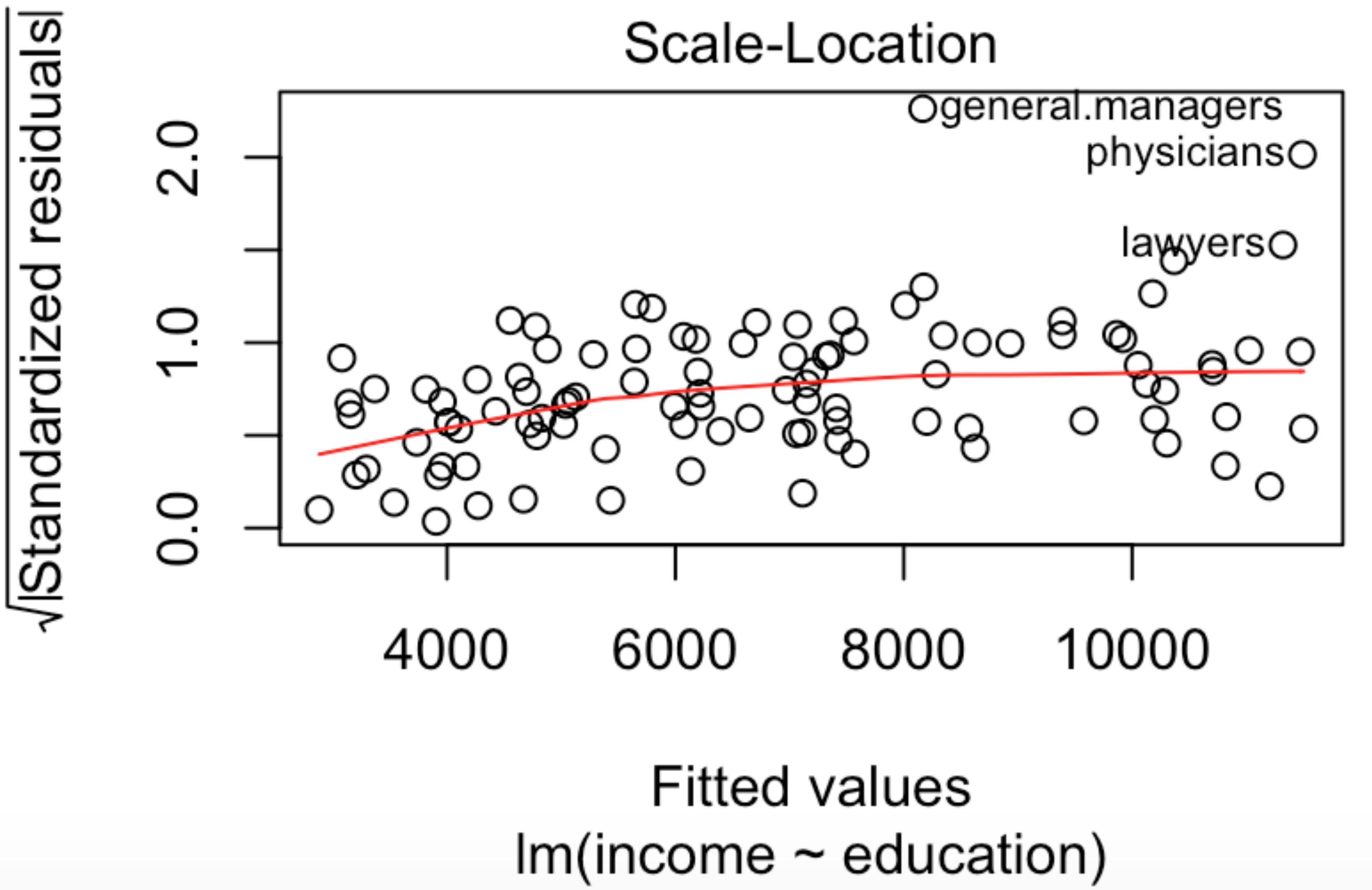
Not Normally Distributed



Linear model: Homoscedasticity (= same variance)

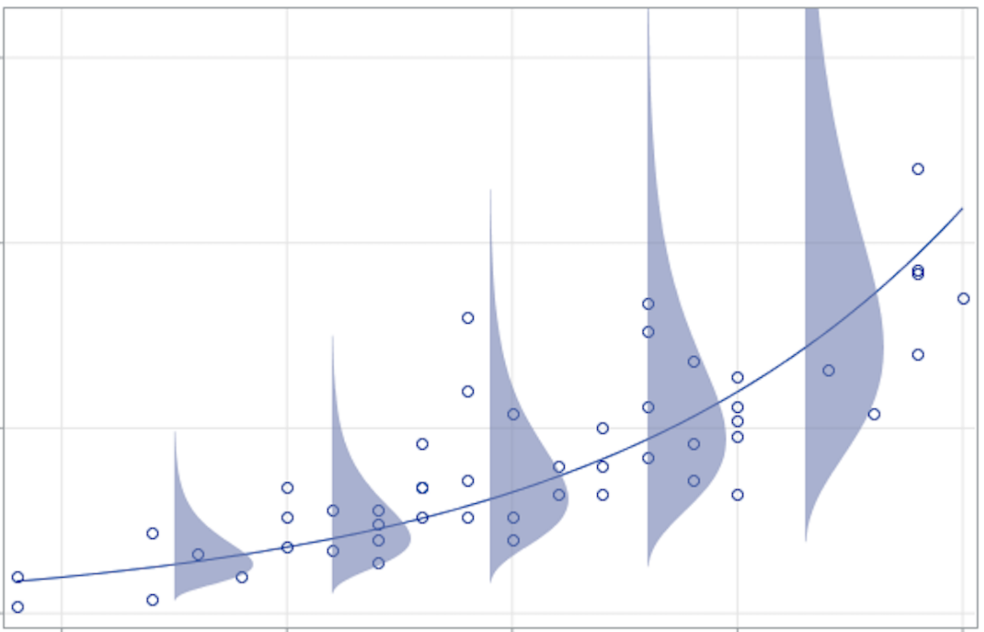
- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity

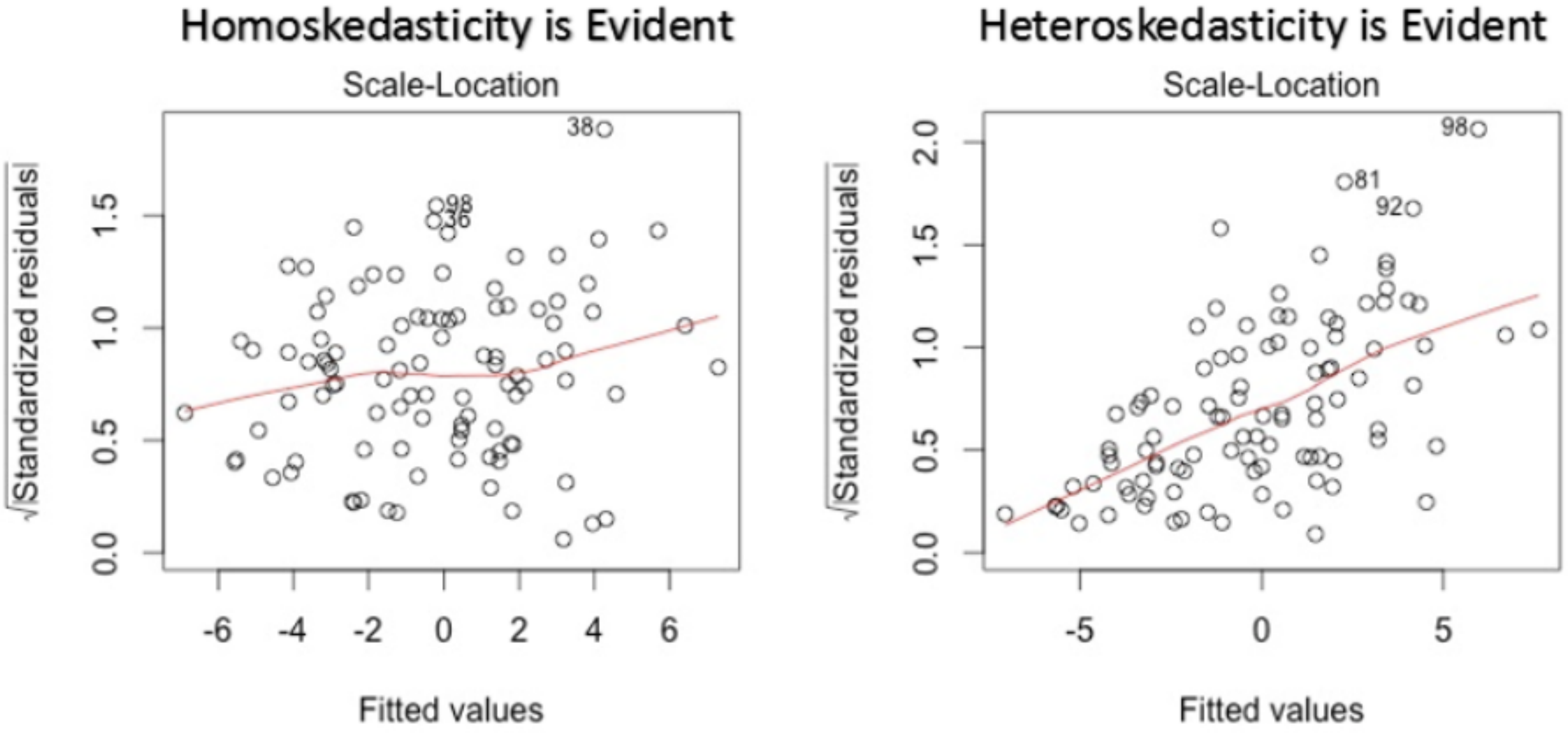




Linear model: Homoscedasticity

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelatio





Linear model: multivariate case

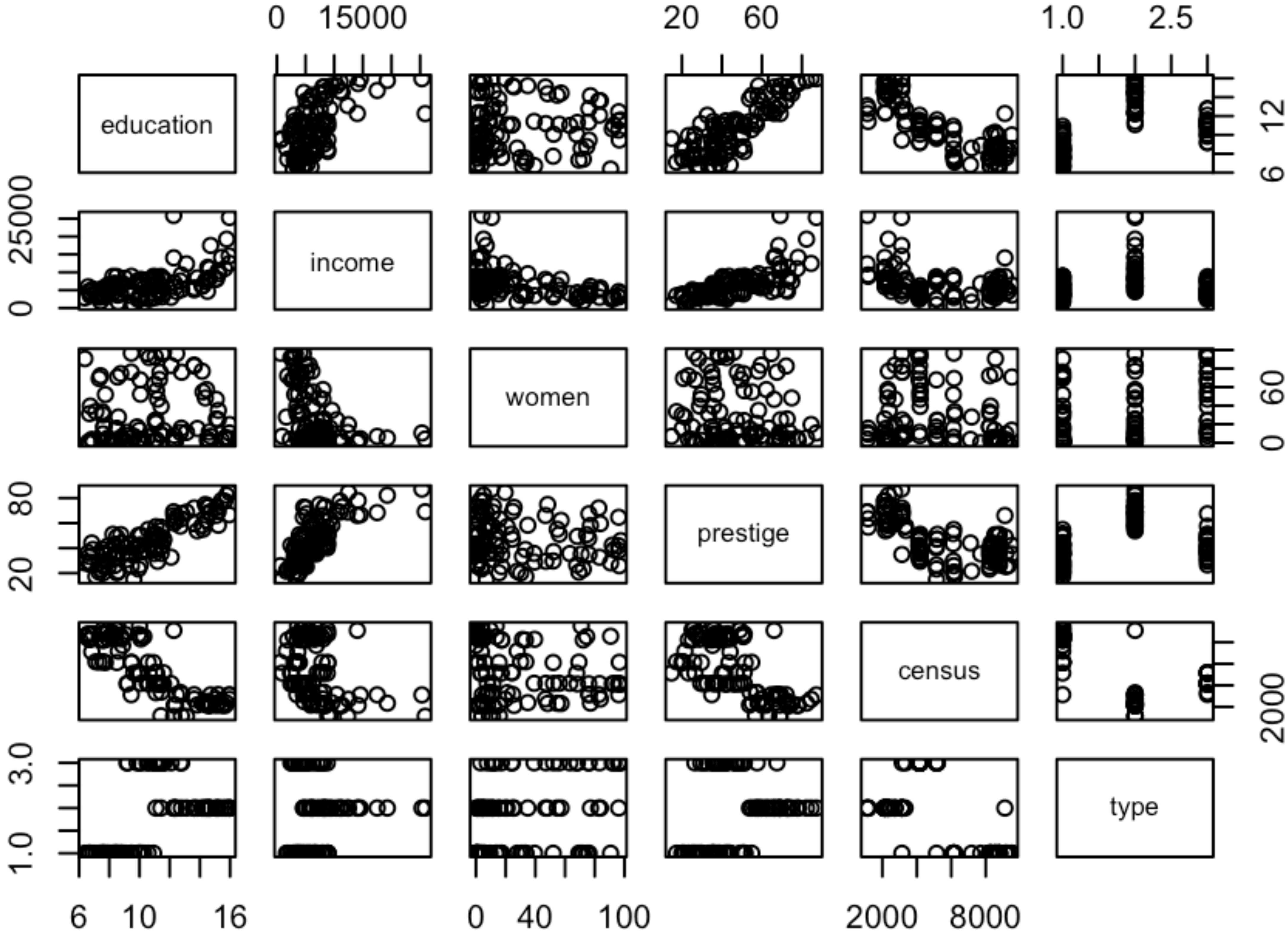
```
Call:
lm(formula = income ~ ., data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-7752.4  -954.6  -331.2   742.6 14301.3 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.32053 3037.27048   0.002  0.99808    
education   131.18372  288.74961   0.454  0.65068    
women       -53.23480    9.83107  -5.415 4.96e-07 ***
prestige    139.20912   36.40239   3.824  0.00024 ***
census        0.04209    0.23568   0.179  0.85865    
typeprof    509.15150 1798.87914   0.283  0.77779    
typewc      347.99010 1173.89384   0.296  0.76757    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2633 on 91 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.6363,	Adjusted R-squared:  0.6123 
F-statistic: 26.54 on 6 and 91 DF,  p-value: < 2.2e-16
```

Linear model: multivariate case



Linear model: multivariate case

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.32053 3037.27048   0.002  0.99808
education    131.18372  288.74961   0.454  0.65068
women        -53.23480    9.83107  -5.415 4.96e-07 ***
prestige     139.20912   36.40239   3.824  0.00024 ***
census         0.04209    0.23568   0.179  0.85865
typeprof     509.15150 1798.87914   0.283  0.77779
typewc       347.99010 1173.89384   0.296  0.76757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Intercept?
- Estimate?
- Std Error? Std Deviation?
- t value?
- P value?

Linear model: multivariate case

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation

- Multicollinearity = predictors that are correlated
- Inflate standard errors
- Warning signs:
  - Regression coefficients change a lot
  - Pairwise correlations
- Solution: VIF (Variance inflation factors), correlation plot

Linear model

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation

1. Regress the Kth predictor on rest of the predictors in the model
2. Compute the R_{sq}

$$\mathrm{VIF} = \frac{1}{1 - R_{sq}}$$

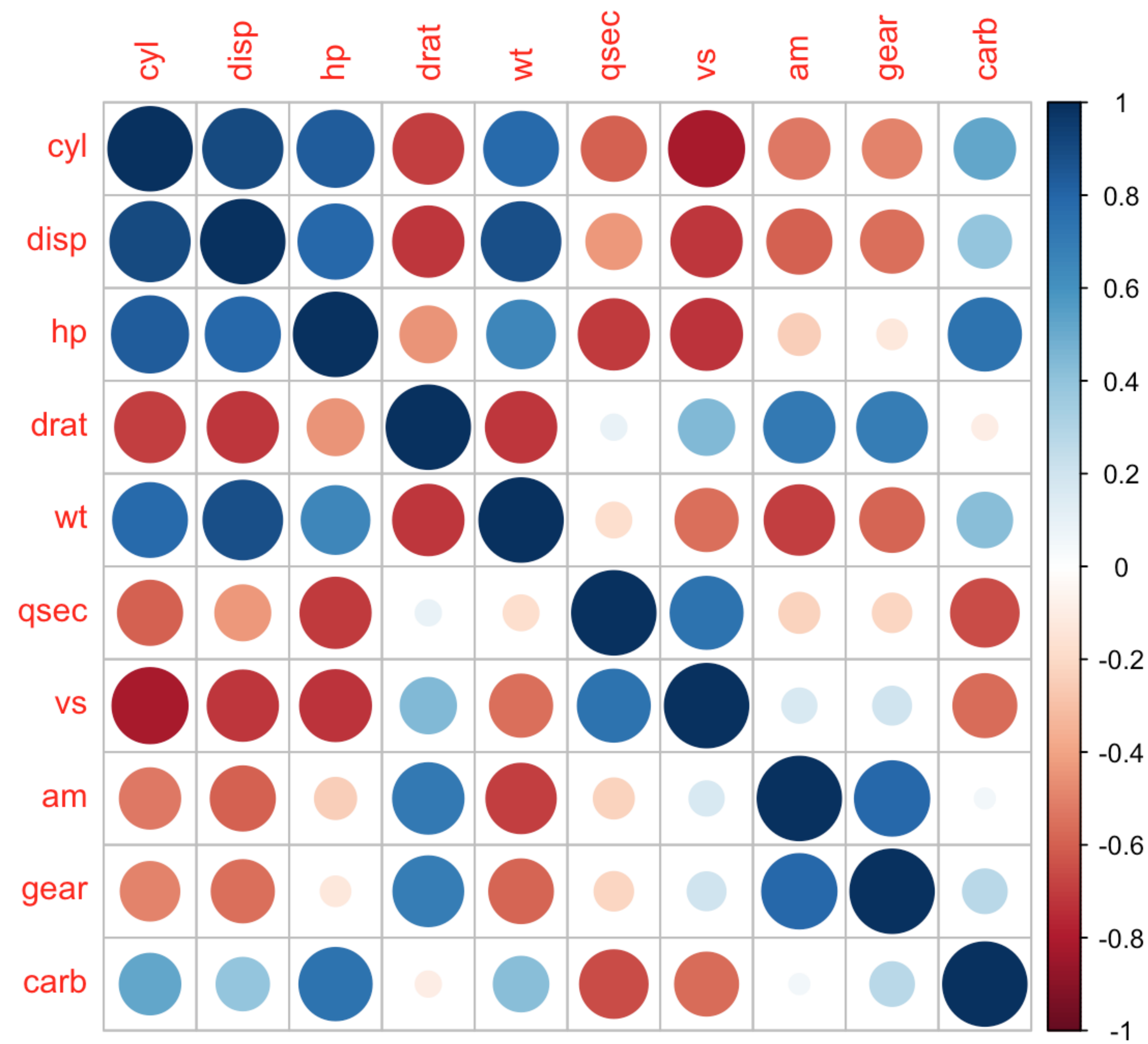
```
library(car)
mod2 <- lm(mpg ~ ., data=mtcars)
vif(mod2)
```

```
> vif(mod2)
      cyl      disp        hp      drat        wt      qsec        vs        am      gear 
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487  5.357452 
     carb 
 7.908747 
```

Linear model

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation

```r
library(corrplot)
corrplot(cor(mtcars[, -1]))
```



Linear model

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation

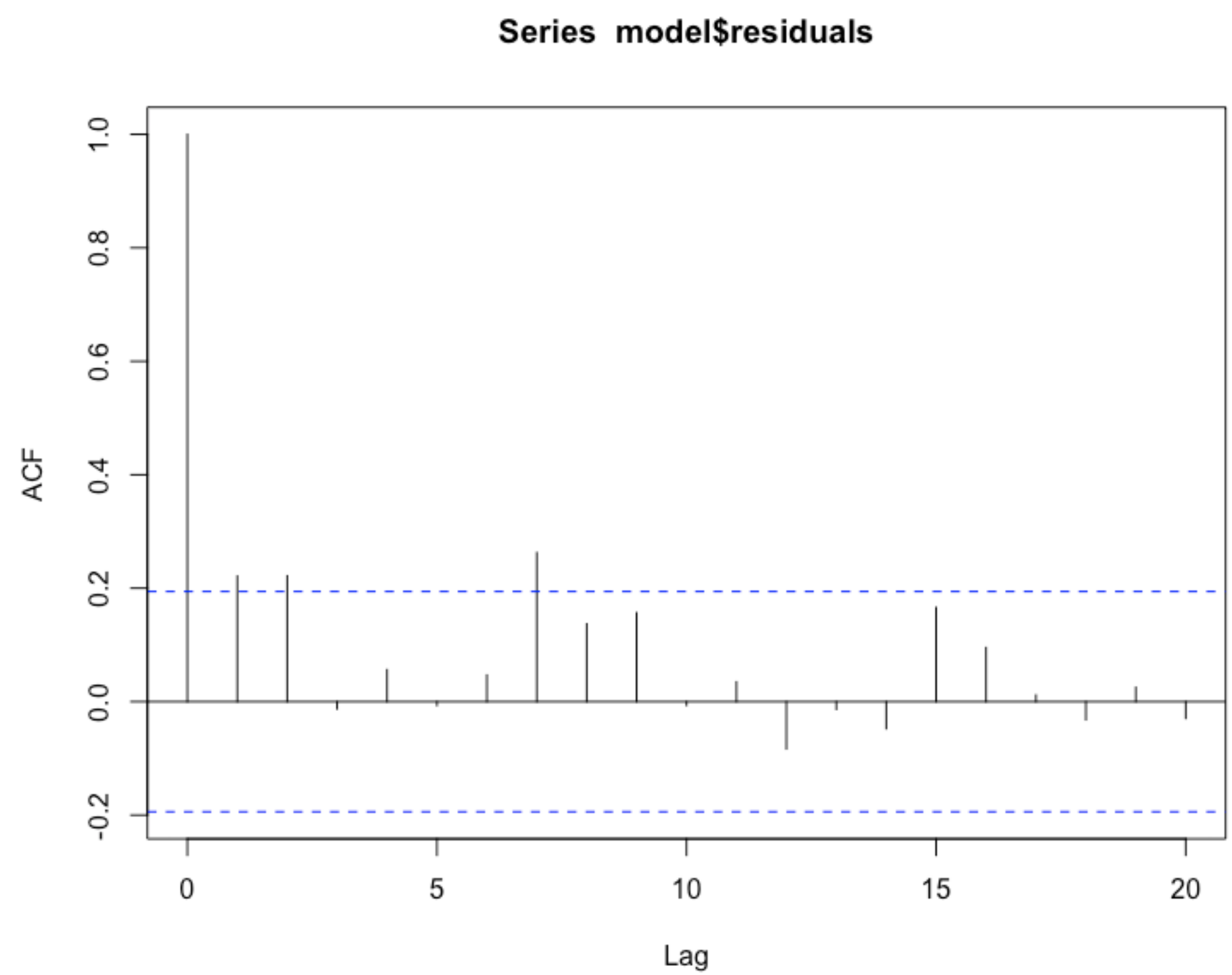
Autocorrelation could be check using acf() plot

- Time serie regression
- Mixed models

Linear model

```
model<-lm(formula = income~education,
   data = newdata)
acf(model$residuals)
```

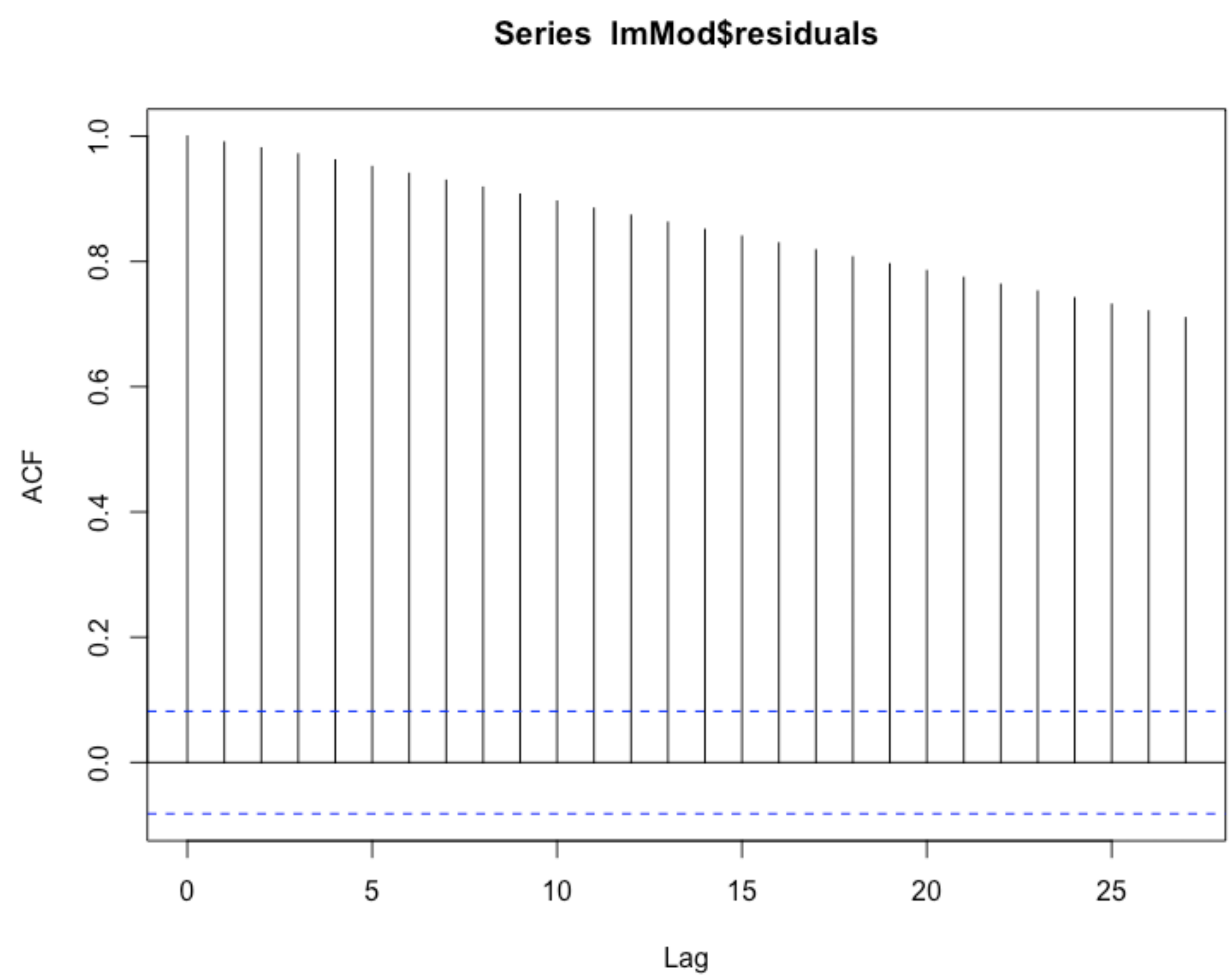
- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation



Linear model

Clearly autocorrelated!

- Linear relationship
- Multivariate normality
- Homoscedasticity
- No or little multicollinearity
- No autocorrelation



Generalized Linear Model

- GLM are extensions of linear models that allow the dependant variable to be non-normal
- It allows the error term to follow another distribution

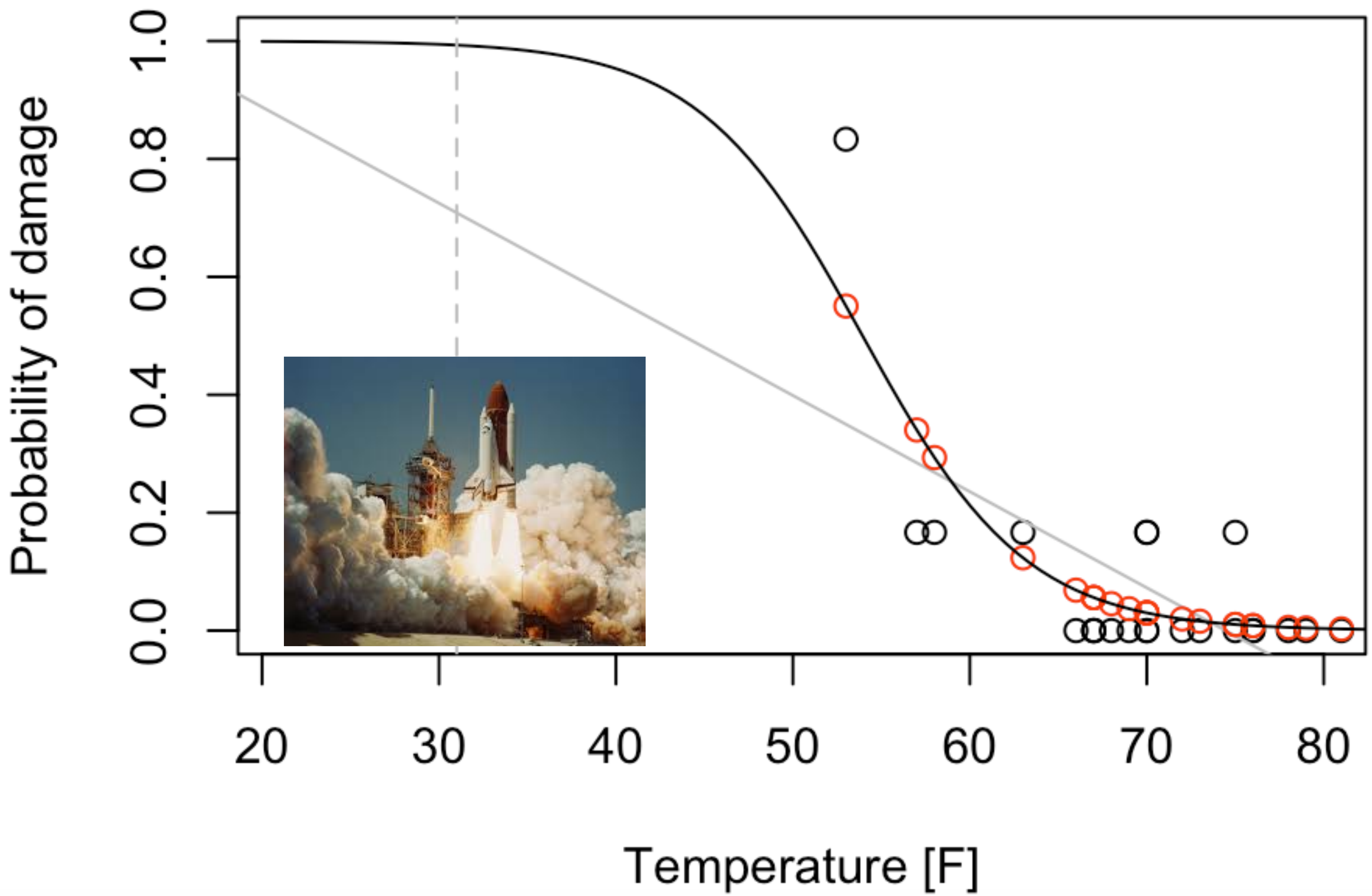
Challenger dataset:

- Estimate probability of defect in function of temperature
- Logistic regression

$$p = P(\text{defect}) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

$$g(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x$$

Generalized Linear Model



Generalized Linear Model

A GLM is made of three components:

- A random component:
  - Conditional distribution of **Y** given the explanatory variables **X**

$$\mathbf{E}[y_i \mid \mathbf{X_i}] = \mu_\mathbf{i}$$

- The systematic component: a linear function of the predictors called the linear predictor

$$\eta = \mathbf{X}\beta \text{ or } \eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- An invertible link function:

$$g(\mu_i) = \eta_i = \mathbf{X}_i^T \beta$$

$$g^{-1}(\eta_i) = \mu_i$$

GLM: link functions

Link name	Function: $\eta_i = g(\mu_i)$	Inverse: $\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
square-root	$\sqrt{\mu_i}$	η_i^2
log	$\log_e(\mu_i)$	$\exp(\eta_i)$
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+\exp(-\eta_i)}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
comp. log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

Standard transformations used with traditiona linear models

Binomial data

GLM:

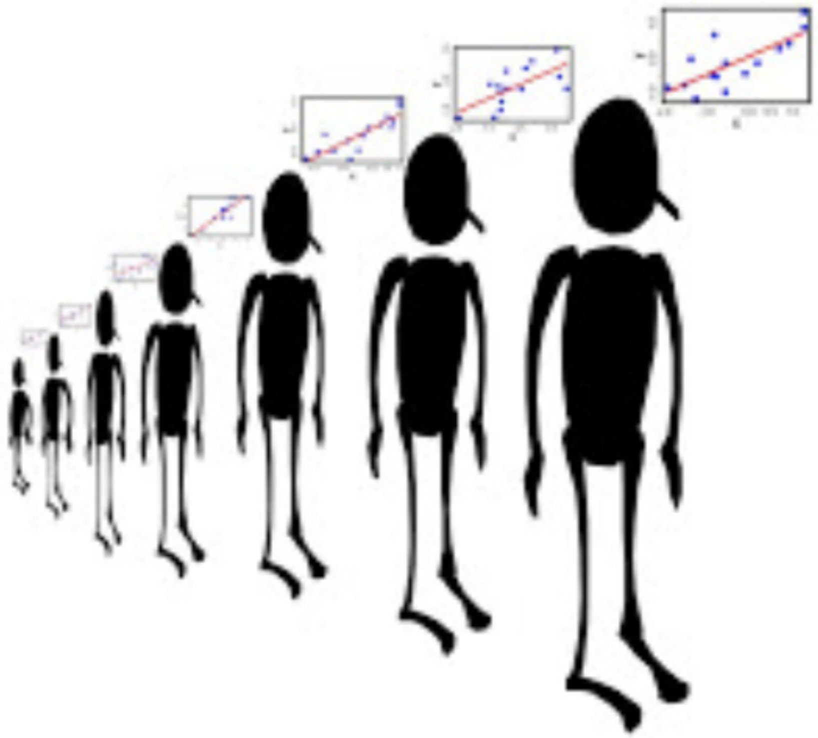
- For every type of distribution, a default canonical link function exists
- Relationship between mean and variance

Family	Notation	Canonical link	Range of y	Variance function, $\mathcal{V}(\mu \mid \eta)$
Gaussian	$N(\mu, \sigma^2)$	identity: μ	$(-\infty, +\infty)$	ϕ
Poisson	$\text{Pois}(\mu)$	$\log_e(\mu)$	$0, 1, \ldots, \infty$	μ
Negative-Binomial	$\text{NBin}(\mu, \theta)$	$\log_e(\mu)$	$0, 1, \ldots, \infty$	$\mu + \mu^2/\theta$
Binomial	$\text{Bin}(n, \mu)/n$	$\text{logit}(\mu)$	$\{0, 1, \ldots, n\}/n$	$\mu(1-\mu)/n$
Gamma	$G(\mu, \nu)$	μ^{-1}	$(0, +\infty)$	$\phi\mu^2$
Inverse-Gaussian	$IG(\mu, \nu)$	μ^2	$(0, +\infty)$	$\phi\mu^3$

Clustering?

Assumptions of linear model violated if:

- several measurements from clusters
- several measurements from the same individuals over time



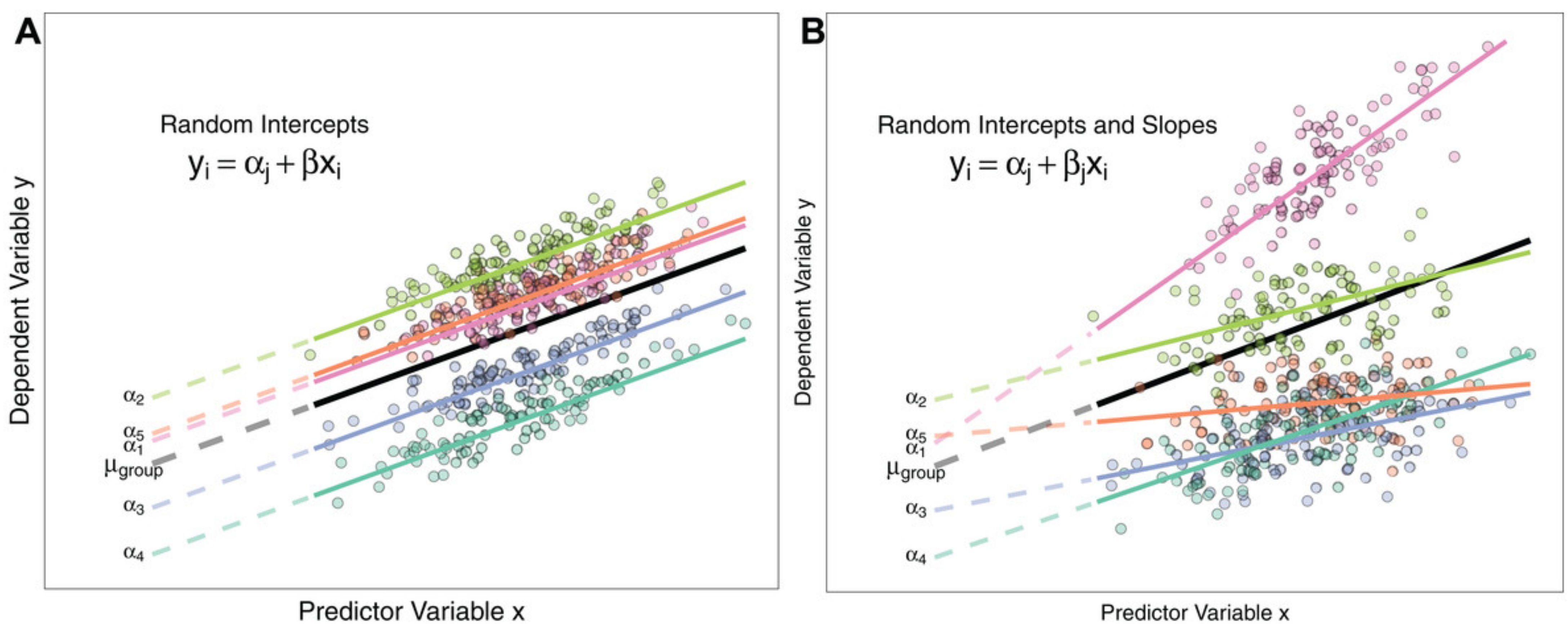
GLM versus GLMM

Linear model

$$y_j = \alpha + \beta x_j + \epsilon_j$$

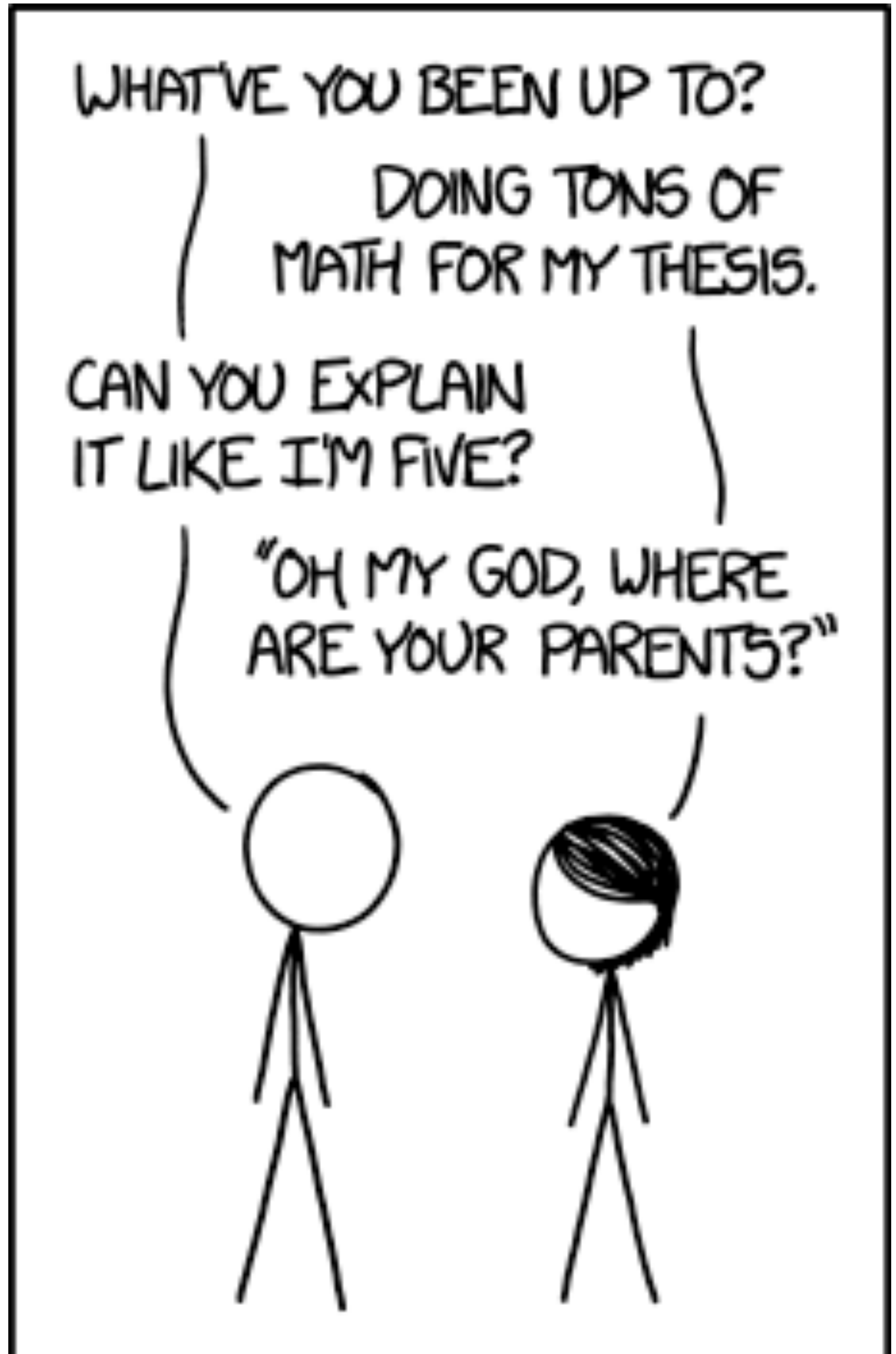
Random intercept linear mixed model

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, \text{ with } \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$



Harrison et al. (2018)

Thank you for your attention



xkcd.com